Chapter 5: Consistency and Limiting Distributions

5.1. Consistency (i.e., Convergence in Probability)

**Definition 1** Let $X_n$ be a sequence of random variables and let $X$ be a random variable. We say that $X_n$ converges in probability to $X$ if for any $\epsilon > 0$,

$$
\lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0 \iff \lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1.
$$

It is expressed by

$$
X_n \xrightarrow{P} X.
$$

In most cases, we choose $X$ as constant (e.g., $c$), leading to

$$
X_n \xrightarrow{P} c.
$$

To show this, we can use the Chebyshev inequality.

**Theorem 1** (Weak Law of Large Number (WLLN)). If $X_1, \cdots, X_n$ are iid random variables with common mean $\mu$ and variance $\sigma^2$, then

$$
\bar{X} \xrightarrow{P} \mu.
$$

**Proof:** By Shebyshev inequality, we have

$$
P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{V(\bar{X})}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} \to 0
$$
as $n \to \infty$.

**Theorem 2** Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then, $X_n + Y_n \xrightarrow{P} X + Y$.

**Proof:** Note that

$$
|X_n + Y_n - (X + Y)| = |(X_n - X) + (Y_n - Y)| \leq |X_n - X| + |Y_n - Y|.
$$

Then, we have

$$
\{|X_n - X| \geq \epsilon/2\} \cup \{|Y_n - Y| \geq \epsilon/2\} \Rightarrow \{|X_n + Y_n - (X + Y)| \geq \epsilon\}.
$$

Then,

$$
P(|X_n + Y_n - (X + Y)| \geq \epsilon) \leq P(|X_n - X| \geq \epsilon/2) + P(|Y_n - Y| \geq \epsilon/2) \to 0
$$
as $n \to \infty$. We draw the conclusion.

**Theorem 3** If $X_n \xrightarrow{P} X$, then $aX_n \xrightarrow{P} aX$ for any constant $a$.

**Proof:** This can be easily shown by $|aX_n - aX| = |a||X_n - X|$.

**Theorem 4** (Continuous Mapping) If $X_n \xrightarrow{P} c$ and $g(\cdot)$ is a continuous function, where $c$ is a constant, then $g(X_n) \xrightarrow{P} g(c)$.
Proof: For any \( \epsilon > 0 \), we find \( \delta > 0 \) such that \( |g(x) - g(c)| \leq \epsilon \) when \( |x - c| \leq \delta \). This is equivalent to
\[
|x - c| \leq \delta \Rightarrow |g(x) - g(c)| \leq \epsilon,
\]
which means
\[
\{ |x - c| \leq \delta \} \subseteq \{ |g(x) - g(c)| \leq \epsilon \}.
\]
Thus,
\[
P(|g(X_n) - g(c)| \leq \epsilon) \geq P(|X_n - c| \leq \delta).
\]
Note that the right side goes to 1 as \( n \to \infty \). We draw the conclusion. \( \Box \)

**Theorem 5** If \( X_n \overset{P}{\to} X \) and \( g(\cdot) \) is a continuous function, where \( c \) is a constant, then \( g(X_n) \overset{P}{\to} g(X) \).

Proof: It is hard. The proof is omitted. \( \Box \)

**Theorem 6** If \( X_n \overset{P}{\to} X \) and \( Y_n \overset{P}{\to} Y \), then \( X_n Y_n \overset{P}{\to} XY \).

Proof: The conclusion is drawn by the continuous mapping theorem. \( \Box \)

**Definition 2** Let \( T = T(X_1, \ldots, X_n) \) be an estimator of \( \theta \). We say \( T \) is a consistent estimator of \( \theta \) if \( T_n \overset{P}{\to} \theta \).

Note: Consistency is a minimum requirement of an estimator. If an estimator is inconsistent, then we cannot use it.

**Example 5.1.1.** Let \( X_1, \ldots, X_n \) be iid random sample from a distribution with mean \( \mu \) and variance \( \sigma^2 \). Then, \( \bar{X} \) is a consistent estimator of \( \mu \). If the fourth moment exists, then \( S^2 \) is a consistent estimator of \( \sigma^2 \).

Solution: By WLLN, we have \( \bar{X} \overset{P}{\to} \mu \). For \( S^2 \), we have
\[
S^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right].
\]

By WLLN, we have
\[
\frac{1}{n} \sum_{i=1}^{n} X_i^2 \overset{P}{\to} \text{E}(X_i^2).
\]

Thus,
\[
S^2 \overset{P}{\to} \text{E}(X_i^2) - \text{E}^2(X_i) = V(X_i) = \sigma^2.
\]

**Example 5.1.2.** Suppose \( X_1, \ldots, X_n \) is an iid sample from \( Uniform[0, \theta] \). Then \( X_{(n)} \) is a consistent estimator of \( \theta \).

Solution:
\[
P(X_{(n)} \leq \theta - \epsilon) = \prod_{i=1}^{n} P(X_i \leq \theta - \epsilon) = \left( \frac{\theta - \epsilon}{\theta} \right)^n \to 0.
\]

Then
\[
P(|X_{(n)} - \theta| \geq \epsilon) = P(\theta - X_{(n)} \geq \epsilon) = P(X_{(n)} \leq \theta - \epsilon) \to 0,
\]

implying the conclusion.

5.2. Convergence in Distribution.
Definition 3 Let \{X_n\} be a sequence of random variables and let \(X\) be a random variable. Let \(F_n\) and \(F\) be the cdf of \(X_n\) and \(X\). Let \(C(F)\) be the set of all continuous points of \(F\). We say that \(X_n\) converges in distribution to \(X\) if

\[
\lim_{n \to \infty} F_n(x) = F(x)
\]

for all \(x \in C(F)\). We denote \(X_n \xrightarrow{D} X\).

Note: If \(X\) is a continuous random variable, then \(F(x)\) is continuous in all \(x \in \mathbb{R}\). Thus, if \(X_n \xrightarrow{D} X\), then \(F_n(x) \to F(x)\) for all \(x \in \mathbb{R}\). This is important in the implementation of the CLT.

The following conclusions are important. We can ignore the proofs.

- Let \(\varphi_n(t)\) be the c.f. of \(X_n\) and \(\varphi(t)\) be the c.f. of \(X\). Then, \(X_n \xrightarrow{D} X\) if and only if \(\lim_{n \to \infty} \varphi_n(t) = \varphi(t)\) for all \(t \in \mathbb{R}\).
- If \(X_n \xrightarrow{P} X\), then \(X_n \xrightarrow{D} X\).
- If \(X_n \xrightarrow{D} b\) for a constant \(b\), then \(X_n \xrightarrow{P} b\).
- If \(X_n \xrightarrow{P} X\) and \(Y_n \xrightarrow{P} Y\), then \(X_n + Y_n \xrightarrow{P} X + Y\).
- If \(X_n \xrightarrow{D} X\), then for any \(g\) continuous function there is \(g(X_n) \xrightarrow{D} g(X)\).
- (Slutsky’s Theorem). Let \(X_n \xrightarrow{D} X\), \(A_n \xrightarrow{P} a\), and \(B_n \xrightarrow{P} b\), then \(A_n + B_nX_n \xrightarrow{D} a + bX\).
- (Lebesgue Dominate Convergence Theorem) Suppose \(\lim_{n \to \infty} f_n(x) = f(x)\) for all \(x\). If \(|f_n(x)| \leq g(x)\) and \(g(x)\) is integrable, then

\[
\lim_{n \to \infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx
\]

for any \(-\infty \leq a < b \leq \infty\).
- Let \(f_n(x)\) and \(f(x)\) be the PDF or PMF of \(X_n\) and \(X\) respectively. If \(|f_n(x)| < g(x)\) for an integrable function \(g\) and \(f_n(x) \to f(x)\) pointwise, then \(X_n \xrightarrow{D} X\).

Example. If \(X_n \sim t_n\), then \(X_n \xrightarrow{D} N(0, 1)\).

Solution: Let \(Y \sim N(0, 1)\) and \(Z_n \sim \chi_n^2\) be independent. Then, the distribution of \(Y/\sqrt{Z_n/n}\) is \(t_n\).

For the denominator, by \(E(Z_n) = n\) and \(V(Z_n) = 2n\). We have \(E(Z_n/n) = 1\) and \(V(Z_n/n) = 2/n\). Thus, \(Z_n/n \xrightarrow{P} 1\), implying that \(Y/\sqrt{Z_n/n} \xrightarrow{D} Y\). Thus, \(X_n \xrightarrow{P} N(0, 1)\).

5.2.1 Bounded in Probability

Definition 4 (Bounded in probability). We say that \(\{X_n; n = 1, 2, \cdots\}\) is bounded in probability, if for all \(\epsilon > 0\), there is \(B_\epsilon > 0\) and an integer \(N_\epsilon\) such that

\[
P(|X_n > B_\epsilon| \geq 1 - \epsilon, \forall \, n > N_\epsilon.
\]

The following notations are often used.

- \(X_n = O_p(1)\) if \(X_n\) is bounded in probability;
• $X_n = O_p(Y_n)$ if $X_n/Y_n$ is bounded in probability;
• $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$;
• $X_n = o_p(Y_n)$ if $X_n/Y_n \xrightarrow{P} 0$.

**Theorem 7** If $X_n \xrightarrow{D} X$, then \{X_n\} is bounded in probability. If $X_n$ is bound in probability and $Y_n \xrightarrow{P} 0$, then $X_nY_n \xrightarrow{P} 0$. If $Y_n = O_p(1)$, and $X_n = o_p(Y_n)$, then $X_n \xrightarrow{P} 0$.

### 5.2.2. $\Delta$-Method

**Theorem 8** Let $X_n$ be a sequence of random variables. If
$$\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$
then for any smooth function $g(x)$ continuous at $\theta$ we have
$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{D} N(0, [\dot{g}(\theta)]^2 \sigma^2).$$

**Proof:** For any fixed $n$, we have the Taylor expansion
$$g(X_n) = g(\theta) + \dot{g}(\theta)(X_n - \theta) + \frac{1}{2}\ddot{g}(\theta^*)(X_n - \theta)^2,$$
where $\theta^*$ is between $\theta$ and $X_n$. Then, we have
$$\sqrt{n}[g(X_n) - g(\theta)] = \dot{g}(\theta)[\sqrt{n}(X_n - \theta)] + \frac{1}{2}\ddot{g}(\theta^*)[\sqrt{n}(X_n - \theta)]^2.$$  
Note that the last term can be ignored. We have
$$\sqrt{n}[g(X_n) - g(\theta)] \approx \dot{g}(\theta)[\sqrt{n}(X_n - \theta)].$$
Since $\dot{g}(\theta)$ is fixed, we draw the conclusion.

**Theorem 9** If $x_n \in \mathbb{R}^m$ is a random vector satisfying
$$\sqrt{n}(x_n - \theta) \xrightarrow{D} N(0, \Sigma),$$
where $\theta \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m \times m}$, then for any function smooth $g(x)$ from $\mathbb{R}^m \to \mathbb{R}^k$ continuous at $\theta$, we have
$$\sqrt{n}[g(x_n) - g(\theta)] \xrightarrow{D} N(0, \dot{g}(\theta)\Sigma\dot{g}^\top(\theta)).$$

**Proof:** By the same method in the proof of the previous theorem, we obtain
$$\sqrt{n}[g(x_n) - g(\theta)] \approx \dot{g}(\theta)[\sqrt{n}(x_n - \theta)] \xrightarrow{D} N(0, \dot{g}(\theta)\Sigma\dot{g}^\top(\theta)).$$
We draw the conclusion. $\diamondsuit$

### 5.3. Central Limit Theorem (CLT)

**Theorem 10** (Central Limit Theorem (CLT)). If $X_1, \cdots, X_n$ are iid sample with finite variance, then
$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$$
where $\mu = E(X_i)$ and $\sigma^2 = V(X_i)$. 

\[ \]
Proof. We use the limit of characteristic function. Let \( \phi(t) \) be the characteristic function of \( X_i \). Without the loss of generality, we can assume \( \mu = 0 \). By the property of \( \phi(t) = E(e^{itX}) \), we have \( \dot{\phi}(0) = 0 \) and \( \ddot{\phi}(0) = \sigma^2 \). This can be shown easily. By the definition of \( \phi(t) \), we have

\[
\frac{\partial}{\partial t} \phi(t) = E\left( \frac{\partial}{\partial t} e^{itX} \right) = E(iX e^{itX}) = it \phi(t).
\]

Let \( t = 0 \). We have \( \dot{\phi}(t) = iE(X) = i \mu \). Similar, we can show that \( \dot{\phi}(t) = i^2 E(X^2) = -E(X^2) = -(\mu^2 + \sigma^2) \). Next, we study the characteristic function of \( \sqrt{n}(\overline{X} - \mu) \), denoted by \( \phi_n(t) \). We have

\[
\phi_n(t) = E(e^{it\sqrt{n}\overline{X}}) = E(e^{it/\sqrt{n} \sum_{i=1}^{n} X_i}) = \prod_{i=1}^{n} E\left(e^{it/\sqrt{n} X_i}\right) = \phi^n(t/\sqrt{n}).
\]

Expand this at 0, we have

\[
\phi_n(t) = \left[ \phi\left(\frac{t}{\sqrt{n}}\right) \right]^n = \left[ 1 + \phi(0) \frac{it}{\sqrt{n}} + \frac{1}{2} \ddot{\phi}(0) \left( \frac{it}{\sqrt{n}} \right)^2 + \frac{1}{6} \phi^{(3)}(0) \left( \frac{it}{\sqrt{n}} \right)^3 \right]^n 
\approx \left[ 1 - \frac{1}{2} \ddot{\phi}(0) \left( \frac{t^2}{n} \right) \right]^n 
\rightarrow e^{-t^2 \sigma^2 / 2}
\]

where \( t^* \) is between 0 and \( t/n \). Since the limit is the characteristic function of \( N(0, \sigma^2) \). We obtain \( \sqrt{n} \overline{X} \xrightarrow{D} N(0, \sigma^2) \) if \( \mu = 0 \). Then, we draw the conclusion easily.

\[\Box\]

Example. (Binomial distribution). Suppose \( X \sim Bin(n, p) \). Let \( \hat{p} = X/n \). Then, \( \sqrt{n}(\hat{p} - p) \xrightarrow{D} N[0, p(1-p)] \). By this property, we can provide \((1 - \alpha)\)-level confidence interval for \( p \) as

\[
\hat{p} \pm z_\alpha \sqrt{\frac{p(1-p)}{n}}.
\]

For any smooth function, we have

\[
\sqrt{n}[g(\hat{p}) - g(p)] \xrightarrow{D} N\{0, [g(p)]^2 p(1-p)\}.
\]

A variance stabilizing transformation is given by \( g \) such that \( [g(p)]^2 p(1-p) = 1 \). We need

\[
g(p) = \frac{1}{\sqrt{p(1-p)}} \Rightarrow g(p) = 2 \arcsin \sqrt{p}.
\]

This can provide another formula for the confidence interval for \( p \).

Example. (Poisson distribution). Suppose \( X_1, \ldots, X_n \sim iid Poisson(\lambda) \). Then, we have \( \sqrt{n}(\overline{X} - \lambda) \xrightarrow{D} N(0, \lambda) \). For any smooth function, we have

\[
\sqrt{n}[g(\overline{X}) - g(\lambda)] \xrightarrow{D} N\{0, [g(\lambda)]^2 \lambda\}.
\]

A variance stabilizing transformation is derived by

\[
g(\lambda) = \frac{1}{\sqrt{\lambda}} \Rightarrow g(\lambda) = 2 \sqrt{\lambda}.
\]
**Example** (Gamma distribution). \( X \sim \Gamma(\alpha, \beta) \) when \( \alpha \to \infty \). Here, we use the PDF as

\[
f(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}.
\]

We have \( \mu = \alpha/\beta \) and \( \sigma^2 = \alpha/\beta^2 \). Thus,

\[
\sqrt{n}(\bar{X} - \frac{\alpha}{\beta}) \xrightarrow{D} N(0, \frac{\alpha}{\beta^2}).
\]

**Example.** (Wald’s method). Let \( \hat{\theta} \) be a consistent estimator of \( \theta \), and \( s_\hat{\theta} \) be its standard error (i.e., \( s_\hat{\theta}^2 \) is an estimator of \( V(\hat{\theta}) \)). Then, the \( (1 - \alpha) \)-level confidence interval for \( \theta \) is

\[
\hat{\theta} \pm z_\alpha s_\hat{\theta}.
\]

We can use the result of confidence interval to test

\[
H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0.
\]

We reject \( H_0 \) iff \( \theta_0 \) does not belong to the confidence interval at significance level \( \alpha \).